Title: Pseudogenization of *Mc1r* gene associated with transcriptional changes related to melanogensis explains leucistic phenotypes in *Oreonectes* cavefish (Cypriniformes, Nemacheilidae)

Zhijin Liu[1,2,†,*], Huamei Wen[1,†], Frank Hailer[3,†], Fang Dong[2,4,†], Zuomin Yang[5], Tao Liu[1], Ling Han[1], Fanglei Shi[2], Yibo Hu[2,*], Jiang Zhou[1, 2,*]

[1] Guizhou Normal University / State Key Laboratory Cultivation for Karst Mountain Ecology Environment of Guizhou Province, Guiyang, Guizhou, 550001, China.

[2] CAS Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chaoyang District, Beijing, 100101, China.

[3] School of Biosciences, Cardiff University, Sir Martin Evans Building, Museum Avenue, Cardiff, CF10 3AX, United Kingdom.

[4] Institute of Physical Science and Information Technology, Anhui University, Hefei, Anhui, 230601, China.

[5] School of Life Sciences, Qufu Normal University, Qufu, Shandong, 273165, China.

† Contributed equally

* Correspondence: Jiang Zhou, zhoujiang@ioz.ac.cn; Yibo Hu, huyb@ioz.ac.cn; Zhijin Liu, liuzj@ioz.ac.cn

Short running title: Leucistic *Oreonectes* cavefish

Abstract

Organisms that have colonized underground caves encounter vastly different selective pressures than their relatives in above-ground habitats. While disruption of certain pigmentation genes has been documented in various cave-dwelling taxa, little is known about wider impacts across pigmentation and other gene pathways. We here study the timeframe and transcriptional landscape of a leucistic and blind cypriniform fish (*Oreonectes daqikongensis*, Nemacheilidae) that inhabits karst caves in Guizhou, China. Based on data from the mitochondrial *ND4*, *ND5* and *Cytb* genes we show that the divergence between *O. daqikongensis* and its most closely related pigmented species occurred ca. 6.82 million years ago (95% HPD, 5.12–9.01), providing ample time for widespread phenotypic change. Indeed, we found that the DNA sequence of *Mc1r* (melanocortin-1 receptor), a key gene regulating the biosynthesis of melanin in most vertebrates, is pseudogenized in *O. daqikongensis*, caused by a 29-bp deletion in the protein-coding region. Furthermore, 99,305 unigenes were annotated based on the transcriptome of skin tissue of *Oreonectes* fish. Among the differentially expressed unigenes, 7,326 (7.4% of the total unigenes) had decreased expression and 2,530 (2.5% of the total unigenes) had increased expression in *O. daqikongensis* skin. As predicted, the expression of *Mc1r* and 18 additional genes associated with melanin biosynthesis were significantly down-regulated in the skin tissue of *O. daqikongensis*, but not in its congener. Our results, integrating with other studies on cavefishes, suggest that loss of pigmentation was caused by coding region loss-of-function mutations along with widespread transcriptional changes, resulting from extended evolutionary time as a cave-dwelling form.

Keywords: *Mc1r*, *Oreonectes daqikongensis*, cavefish, frameshift mutation, leucism, transcriptome.

Introduction

Uncovering the speciation process and genetic basis of phenotypic adaptations of animals to a specific environment is a key goal in evolutionary and comparative zoology. Underground rivers and caves are a globally widespread ecosystem that have been independently colonized by numerous animals (McGaugh et al., 2014; Bilandžija et al., 2013). Since light is lacking year-round, many animals inhabiting such habitats show degeneration of skin pigmentation (Gross & Wilkens, 2013). However, the timeframes on which cave dwellers have evolved their specialized phenotypes, and the molecular mechanisms underlying such phenotypes, remain largely unknown.

The melanocortin-1 receptor (MC1R) gene plays a key role in the regulation of melanin biosynthesis (Majerus & Mundy, 2003). The MC1R receptor protein encoded by *Mc1r* gene belongs to the G-protein coupled receptors family, which has seven transmembrane α-helices. α-MSH (α-Melanocyte-Stimulating hormone) binds to and activates MC1R receptor at the membrane of melanophore cells and promotes the biosynthesis of cAMP (cyclic adenosine monophosphate), followed by the synthesis of melanin (Barsh & Cotsarelis, 2007). High levels of basal *Mc1r* signaling cause increased expression of the microphthalmia-associated transcription factor (MITF), tyrosinase (TYR), tyrosinase-related protein 1 (TYRP1) and other melanogenesis-related genes, which leads to increased eumelanin synthesis (Nishimura, 2011).

Mc1r has been identified as the key candidate gene for animal pigmentation, and several *Mc1r* mutations have been shown to explain the presence of leucistic/light-colored phenotypes (Majerus & Mundy, 2003). Leucism describes a partial loss of pigmentation, with phenotypes such as white or patchily white-colored skin, hair or

70 feathers, but pigmentation still being partially present, e.g. the pigment cells in the
71 eyes which are not affected by the leucistic condition. Albinism, on the other hand, is
72 a phenotype with complete absence of melanin, resulting in albinistic skin,
73 feathers/hair and eyes. Vertebrates with albinism are not only white (or sometimes
74 pale yellowish) in exterior coloration, but also possess pale eyes, often pink or red in
75 color due to blood vessels showing through. In mammals, the leucistic phenotype of
76 *Ursus americanus kermodei* is caused by the substitution of tyrosine by cysteine at the
77 298th amino acid (Y298C) of MC1R (Ritland, Newton & Marshall, 2001). The
78 substitution of arginine by cysteine at 18th amino acid and arginine by tryptophan at
79 the 109th amino acid (R18C and R109W) of MC1R also result in the light-coated
80 phenotype of *Chaetodipus intermedius* (Nachman, Hoekstra & D'Agostino, 2003). In
81 reptiles, the light-colored skins are also associated with mutations of *Mc1r* in lizards
82 and geckos (Mendes et al., 2018; Machado et al., 2019). In birds, the leucistic
83 phenotype of the plumage in *Falco rusticolus* is caused by the substitution of valine
84 by isoleucine at the 128th amino acid of MC1R (Zhan et al., 2012; Johnson, Ambers &
85 Burnham, 2012). Furthermore, melanistic plumage polymorphism in lesser snow
86 geese (*Anser caerulescens*) and arctic skuas (*Stercorarius parasiticus*) is tightly
87 associated with sequence variation within the *Mc1r* gene (Mundy et al., 2004). In fish,
88 two different *Mc1r* alleles (963 and 969 bp long) were found in wild populations of
89 guppies (*Poecilia reticulata*) and individuals with the 963 bp *Mc1r* allele tended to
90 show less black pigmentation than those with the 969 bp allele (Ayumi et al., 2011).

91 The genus *Oreonectes* (Cypriniformes, Nemacheilidae) was first described by
92 Günther (1868), with *O. platycephalus* as the type species, and a total of 18 species of
93 *Oreonectes* are currently considered as valid (Deng et al., 2016; Günther, 1868).
94 Besides *O. platycephalus*, *Oreonectes* fish are semi-cave-dwelling or cave-restricted

95 species, rendering the genus an interesting model to study evolutionary consequences
96 ocolonization of cave habitats (Deng et al., 2016). One particularly specialized species
97 is *O. daqikongensis*, which was only recently described (Deng et al., 2016). The
98 species exhibit numerous typical characteristics of cave dwellers (e.g., lack of eyes),
99 but is the only leucistic *Oreonectes* fish which lack pigmentation (Fig.1a and b) (Deng
100 et al., 2016). We here ask the following evolutionary questions: what is the timeframe
101 on which *O. daqikongensis* evolved the leucistic phenotype? Is the *Mc1r* gene
102 involved in the occurrence of leucism? More broadly, how is the translational
103 landscape of the melanogenesis pathway changed in this species? To answer these
104 questions, we first reconstructed a dated phylogeny of five *Oreonectes* species based
105 on mitochondrial *Cytb*, *ND4* and *ND5* gene sequences. To determine whether the
106 *Mc1r* gene is functional in *O. daqikongensis*, the locus was sequenced in *O.*
107 *daqikongensis* and another three *Oreonectes* species. Finally, to illuminate the
108 evolutionary impacts of long-term cave-dwelling on *Oreonectes* fishes, determine the
109 function of other melanogenesis-related genes in *O. daqikongensis*, we further
110 collected skin tissue samples of *O. daqikongensis* and its close-related species with
111 pigmentation and performed whole-transcriptome sequencing.

112 ## Materials and methods

113 ### Sample collection

114 Fifteen individuals of *O. daqikongensis*, 16 *O. jiarongensis* individuals, six *O.*
115 *dongliangensis* individuals and ten *O. shuilongensis* individuals were collected in
116 Libo and Sando county of Guizhou Province (Fig. 1). These samples have been
117 deposited in the Museum of Guizhou Normal University (for catalog numbers see
118 Table 1).

Phylogenetic analyses and divergence time estimation

DNA extraction was performed for all the collected samples according to the manufacturer's instructions (Qiagen Tissue Kit). To infer the phylogenetic relationships of *Oreonectes* species, the mitochondrial DNA genes (*ND4*, *ND5* and *Cyt b*) of each one individual of herein studied four *Oreonectes* species were amplified with primer sets for Nemacheilidae mitogenome (Rui et al. 2012; Wang et al. 2016; Table 2). Mitogenome sequences of the other 53 species in Nemacheilidae, two species in Cobitidae and two species in Balitoridae were obtained from GenBank (Supplementary Table S1). Sequences were aligned using ClustalX (Thomps et al., 1997) and rechecked by eye. MEGA v7.0 (Sudhir et al., 2016) was used to estimate genetic distances between the taxa in *Oreonectes* based on P-distance method. The Bayesian Inference (BI) of phylogenetic reconstruction was employed using MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003), using species from Cobitidae (*Cobitis sinensis* and *Misgurnus anguillicaudatus*) and Homalopteridae (*Jinshaia sinensis* and *Sinogastromyzon sichangensis*) as outgroups. MODELTEST 3.06 (Posada & Crandall, 1998) was run to determine the appropriate model of sequence evolution in a likelihood ratio test. For Bayesian phylogenetic inference, four Markov chain Monte Carlo (MCMC) runs were performed for 100,000 generations, sampling every ten generations. The initial 5% of trees were discarded as burn-in.

A Bayesian tree for estimating the divergence time was reconstructed in BEAST 1.61 (Drummond & Rambaut, 2007) using the GTR +I+G model, which provided the best fit. Trees were produced based on two independent runs of 10 million MCMC steps each sampling every 2,000th iteration, discarding the initial 25% of values as burn-in. We utilized the calibrated Yule model, as recommended by Heled &

Drummond (2012) for analysis of sequences from different species. We employed calibration points obtained from the dated Nemacheilidae phylogenetic tree by Wang et al. (2016) to (1) place a log normally distributed prior on the age of the root of the tree containing all samples of *Oreonectes* species and outgroups, based on estimated divergence time between Cobitidae and Nemacheilidae at approximately 30 million years ago (Mya; Frickhinger, 1991), and (2) a prior for the origination of the genus *Cobitis* as 13.8–15.9 Mya (Zhou, 1992). An uncorrelated lognormal model of lineage variation with a constant population size tree prior was employed (Drummond et al., 2012). Convergence of two independent MCMC runs was assessed in Tracer v.1.6 (available at http://tree.bio.ed.ac.uk), as was convergence of model parameter values, i.e. effective sample size [ESS] values being >200. The tree and posterior distribution were summarized with TreeAnnotator v.1.5.3 and visualized by FigTree v.1.4.3 (available at http://tree.bio.ed.ac.uk).

DNA extraction, primer design, polymerase chain reaction and DNA sequencing of *Mc1r* gene

The candidate *Mc1r* gene was amplified by polymerase chain reaction (PCR). Primers used in this study are MC1R-F (5'-GAATATCAGAGGTGTGCTGAAGC-3') and MC1R-R (5'-TCCTTGAGAGTCTTGCGCAG-3') which were designed based on alignments of the flanking regions of *Mc1r* in *Astyanax mexicanus*, *Carassius aumtus* and *Triplophysa rosa*. The *Mc1r* coding region was amplified for two individuals each of *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis*, and three individuals of *O. daqikongensis.* PCR was carried out in 50 μl reaction mixtures containing 50 mM KCl, 10 mM Tris-HCl, 1.5 mM Mg^{2+}, 200 μmol of each dNTP, 0.2 μmol of each primer, 1.5 U Hotstart Taq DNA polymerase (QIAGEN), 1 μg/μL BSA and ≤10 ng of genomic DNA. The PCR cycling conditions were as follows: initial denaturation at

95°C for 5 min, followed by 32 cycles each of denaturation (95°C for 30 sec), annealing (at 58°C for 30 sec), and elongation (at 72°C for 60 sec) with a final elongation step at 72°C for 10 min. The PCR products were examined for specificity following electrophoresis on 1.5% agarose gels, then sequenced on an ABI 3130 xl DNA Sequencer (Applied Biosystems).

DNA sequence analysis of the *Mc1r* gene

DNA sequencing data was read and assembled by SeqMan II (DNASTAR Lasergene, version 6). The open reading frame (ORF) analysis and amino acid sequence inference was completed by using the search ORF program of EditSeq 6.1 (DNASTAR Lasergene, version 6.1). Clustalw2 (http://www.ebi.ac.uk/Tools/msa/clustalo) was applied to perform nucleotide alignment. The protein transmembrane structure was predicted and functionally analyzed using TMHMM v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/), and MODELLER (https://swissmodel.expasy.org/; Eswar et al., 2008) was used to construct the three-dimensional structure of the MC1R protein.

Total RNA isolation, library preparation, and sequencing

Skin tissues were isolated from the rear of three *O. daqikongensis* and three *O. jiarongensis* individuals immediately after capture in the wild. The isolated tissues were immediately put into liquid nitrogen, frozen and stored at -80°C. Total RNA was isolated from each sample of tissue using TRIzol reagent (Life Technologies, CA, US) according to the manufacturer's instructions. The purity and content of each RNA sample were measured using the Qubit RNA Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies, CA, USA) and confirmed by running total RNA samples on 1 % agarose gels.

One paired-end (PE) cDNA library was generated for each sample. Illumina sequencing was conducted at Beijing Novogene Biological Information Technology Co., Ltd., Beijing, China (http://www.novogene.com/) using the Illumina TruSeqTM RNA Sample Preparation Kit (Illumina, San Diego, CA, USA) following the manufacturer's recommendations (Kircher Stenzel & Kelso, 2009). Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under an elevated temperature in the proprietary Illumina fragmentation buffer. First-strand cDNA was synthesized using random oligonucleotides and Super Script II (Life Technologies). Second-strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities, and enzymes were then removed. After adenylation of 3' ends of DNA fragments, Illumina PE adapter oligo nucleotides were ligated to prepare for hybridization. To preferentially select cDNA fragments of 300 bp in length, the library fragments were purified with the AMPure XP system (Beckman Coulter, Beverly, MA, USA). DNA fragments with ligated adaptor molecules on both ends were selectively enriched using Illumina PCR Primer Cocktail in a 10-cycle PCR (polymerase chain reaction). Products were purified (AMPure XP system) and quantified using the Agilent High-Sensitivity DNA assay on the Agilent Bioanalyzer 2100 system (Agilent Technologies Co. Ltd., Beijing, China). Clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq 4000 platform and 2x150 bp paired-end reads were generated.

The assembly of transcriptome data, gene annotation and analysis

Raw sequence data reads in FASTA format were first processed using in-house Perl scripts. In this step, clean reads were obtained by removing adapters, poly-N, and low-quality sequences from the raw read data. To generate a common and nonredundant *Oreonectes* transcriptome database for further RNA-seq analyses, transcriptome assembly was carried out based on clean data from *O. jiarongensis* using Trinity software (v.2.0.6), with all parameters set as default (Grabherr et al., 2011). The resulting transcriptome assembly was trimmed with a custom perl script to remove contigs under 500 base pairs (Maytin et al., 2018). Before annotation, unigenes were picked from the transcriptome with CD-hit (Li & Godzik, 2006). Intactness of the assembled *O. shuilongensis* transcriptome was assessed with the software tool BUSCO (Benchmarking Universal Single-Copy Orthologs) that is based on evolutionarily informed expectations of gene content, with default settings (Simão et al., 2015; Waterhouse et al. 2018). The unigenes were then annotated with BLASTx (BLAST + v.2.2.25) by querying these to the following databases: NCBI non-redundant protein sequences (Nr), NCBI non-redundant nucleotide sequences (Nt), the Protein Family database (Pfam), Swiss-Prot, Gene Ontology (GO), the Eukaryotic Orthologous Groups database (KOG) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). The E-value cutoff was set at 1×10^{-5}.

Gene expression quantification and identification of differentially expressed genes (DEGs)

We mapped trimmed Illumina reads from the skin tissue from *O. daqikongensis* and *O. jiarongensis* to *Oreonectes* transcriptome using Bowtie2 with default parameters (Langmead, 2010). Gene expression levels were estimated by RSEM (v.1.2.15) for each sample (Li & Dewey, 2011). First, the read count for each gene was obtained from the result of clean data mapped back onto the assembled transcriptome

reference. Subsequently, the read count of each sequenced library was adjusted with the edger program package to 'fragments per kb per million mapped (FPKM). The edgeR package (http://bioconductor.org/packages/release/bioc/html/edgeR.html) was used to obtain the "base mean" value for identifying DEGs. A false discovery rate (FDR) ≤0.05 and the absolute value of log2 ratio ≥1 (two-fold change difference) were set as the thresholds for the significance of the gene expression difference between two groups. In addition, information for DEGs was collected from unigenes annotations, and these genes were subjected to GO and KEGG significant enrichment analyses to identify biological functions and metabolic pathways in which these genes participate. Furthermore, we collected information for candidate genes involved in melanogensis.

Results

Phylogenetic analysis and divergence time estimation

To infer the phylogenetic relationships of *Oreonectes* species, the phylogeny of five *Oreonectes* species and other 53 species in Nemacheilidae was reconstructed using two species in Cobitidae and two species in Balitoridae as outgroups. Five mitochondrial fragments were amplified for each one individual of herein studied four *Oreonectes* species and the continuous mitochondrial sequence were aligned, including the complete sequences of *ND4* (1,382 bp), *ND5* (1,837 bp) and *Cytb* (1,141 bp) (Table 2). In the concatenated alignment (*ND4*: 1st -1,382nd bp, *ND5*: 1,383rd - 3,219th bp, *Cytb*: 3,220th -4,360th bp) that comprised 4,360 bp across 61 species, we found 2290 variable sites out of the alignment (Supplementary data 1). The genetic distances among *Oreonectes* species ranged from 0.010 to 0.149 (Table 3). GTR+I+G model was the best fitting model for each of these three mitochondrial genes

individually, and the same substitution model was also identified as the best-fitting model for the concatenated sequence, with a gamma shape parameter of 1.0419, a transition/transversion ratio of 6.23 and base frequencies of A =0.3226, C= 0.2148, G= 0.1366 and T= 0.3260. Phylogenetic analyses resulted in virtually identical tree topologies (Fig. 1d), with a basal divergence among *Oreonectes* species at ca. 17.53 Mya (95% HPD, 13.93–22.13 Mya), and a split between *O. daqikongensis* and its most closely-related species (*O. dongliangesis* and *O. jiarongensis*) at 6.82 Mya (95% HPD, 5.12–9.01 Mya; Fig. 1d).

Pseudogenization of *Mc1r* gene in *O. daqikongensis*

The amplicon length of *Mc1r* gene ranged from 971–1,004 bp. The protein coding region of the *Mc1r* gene of *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis* was 988 bp, while it was only 955bp in *O. daqikongensis* (Supplementary Fig. S1 and Supplementary data 2). The *Mc1r* sequences from 15 *O. daqikongensis* individuals all had a 29 bp deletion (starting at nucleotide position 236 of the alignment), a six bp insertion (starting at nucleotide position 819) and a nine bp deletion (starting at nucleotide position 867). The first 29 bp deletion caused a frame-shift mutation of *Mc1r* gene of *O. daqikongensis*, resulting in a premature STOP codon (TGA; starting at nucleotide position 279-281), while *Mc1r* sequences were intact in *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis* (Fig. 1e). Thus, *Mc1r* is pseudogenized in *O. daqikongensis*, showing truncation of the protein within the intracellular loop. The MC1R receptor protein sequence of *O. daqikongensis* was predicted to be incapable of generating a functional transmembrane domain through TMHMM analysis, while the *Mc1r* receptor proteins of other *Oreonectes* species with pigmented skin (*O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis*) were predicted to contain an intact transmembrane domain (Fig. 2a and 2b). Similarly, the three-dimensional

structure of the *Mc1r* receptor protein predicted by MODELLER exhibited disrupted transmembrane domain in *O. daqikongensis* (Fig. 2c and 2d), but not in the other species.

Sequencing, *de novo* assembly, and annotation of the *Oreonectes* Transcriptome

After trimming adapters and removing low quality reads, 67,824,194 – 44,390,242 pair-end reads were retained for three individuals of *O. daqikongensis* and three individuals of *O. jiarongensis*. Then *de novo* assembly of *O. jiarongensis* transcriptome was performed with all the clean data of three individuals of *O. jiarongensis*. We got 99,305 unigenes and the total length, average length and N50 value of the unigenes are 207,947,183 bp, 2,039 bp, and 3,571 bp, respectively (Fig. 3a). Of the 4,584 Benchmarking Universal Single-Copy Orthologs (BUSCOs) for Actinopterygii, 89.60% were complete and a further 5.14% were fragmented in *O. jiarongensis* transcriptome. It suggested that the *de novo* assembly was relatively successful comparing to several fish transcriptomes assembly (Supplementary Fig. S2 and Supplementary Table S2). We then annotated the unigenes using seven functional databases. Finally, 71,877 (NR: 72.38%), 77,110 (NT: 77.65%), 66,773 (Swissprot: 67.24%), 42,045 (KOG: 42.34%), 57,845 (KEGG: 58.25%), 71,542 (GO: 72.24%), and 63,813 (Pfam: 64.26%) of the unigenes were annotated (Fig. 3b).

Identification of differentially expressed genes (DEGs)

The mapping rate for transcriptome of *O. daqikongensis* and *O. jiarongensis* to *O. jiarongensis* transcriptome assembly was 72.34% ± 3.1×10^{-3} and 75.26% ± 2.8×10^{-3}, respectively, suggesting that the transcriptome was well-assembled. We identified 9,856 unigenes as differentially expressed, based on a >2-fold change between *O. daqikongensis* and *O. jiarongensis* with an FDR corrected *P*-value (*q*-value) less than

0.05 (log2-fold changes [log_2FC]> 1, [-log_{10}*q*-value]>1.3). Among the differentially expressed unigenes, 7,326 (7.4% of the total unigenes) had decreased expression and 2,530 (2.5% of the total unigenes) had increased expression in *O. daqikongensis* skin (Fig. 4a). Nineteen genes belonging to the GO term "melanin biosynthetic process" and KEGG pathway "melanogenesis" (04916) were down-regulated, in which the expression level of twelve genes (including *Mc1r*) were low in *O. daqikongensis* (Fig. 4b and 4c).

Discussion

The divergence between *O. daqikongensis* and its most closely related species with pigmentation (*O. dongliangensis* and *O. jiarongensis*) was estimated to be ~ 6.82 Mya (Fig 1d and Supplementary Fig 1), and our data suggested that the pseudogenization of *Mc1r* occurred in the leucistic *O. daqikongensis* after its divergence from non-leucistic congeners. Given the lack of informative variable sites in the *Mc1r* alignement between *O. daqikongensis* and its closest relative, estimation of the *Mc1r* pseudogenization time is infeasible (Supplementary Fig. S1; Zhao et al., 2010). Nevertheless, the presence of several indels in *O. daqikongensis Mc1r*, along with the widespread and drastic transcriptional changes in this lineage, it appears that *O. daqikongensis* represents a relatively ancient cave-dwelling species in which widespread adaptations and loss-of function mutations have accumulated. This renders our study species an interesting system to understand long-term evolutionary consequences of cave dwelling.

As fish living in caves, *O. daqikongensis*, *A. mexicanus*, and *Sinocyclocheilus anshuiensis* all have the leucistic phenotype (McGaugh et al., 2014; Bilandžija et al., 2013; Gross, Borowsky & Tabin, 2009; Gross & Wilkens, 2013; Jeffery et al., 2015;

Stahl & Gross, 2015; Yang et al., 2016). Similar selective pressures can lead to the parallel evolution of identical or similar traits in distantly related species, often referred to as adaptive phenotypic convergence (Christin, Weinreich & Besnard, 2010; Stern, 2013; Storz, 2016). A critical mechanism underlying phenotypic convergence is genetic convergence, including the same metabolic and regulatory pathways, protein-coding genes, or even identical amino acid substitutions in the same gene (Christin, Weinreich & Besnard, 2010; Stern, 2013; Storz, 2016; Mundy, 2005). Compared to the cavefish *A. mexicanus*, the mechanism of leucism in *Oreonectes* cavefish was also caused by pseodugenized *Mc1r*, representing a remarkable scenario of genetic convergence (Gross, Borowsky & Tabin, 2009). The pseudogenization of *Mc1r* in *O. daqikongensis* could directly or indirectly led to the low-level expression of dozens of melanogenesis related genes. As a result, the biosynthesis of melanin is blocked, leading to the albino phenotype.

In *A. mexicanus*, the reduced melanin phenotype of *A. mexicanus* is also due to the *Oca2* gene mutation, and the molecular mechanism and the degree of degeneration in skin pigmentation vary in different populations and circumstances. In Molino cave, *A. mexicanus* lost the 21st exon of *Oca2*, while in Pachón cave it was the deletion of exon 24-3'UTR which resulted in leucism (Protas et al., 2006). In the case of *S. anshuiensis*, genome analysis discovered that a mutation of glycine to arginine (G420R) in exon 4 of the *Tyr* gene encoding tyrosinase might result in leucism of the skin (Yang et al., 2016). Tyrosinase is the rate-limiting oxidase for controlling the production of melanin. It can catalyze the production of melanin and other pigments from tyrosine by oxidation (Kumar et al., 2011). Furthermore, a deletion of *Mpv17* might have also played a role in the leucism of *S. anshuiensis.* In the case of *O. daqikongensis*, besides the genes directly regulated by MC1R, other melanogenesis-

364 related genes involved in the Wnt signaling pathway and MAPK signaling pathway
365 were also down-regulated (Fig. 4b). Our results, integrating with above-mentioned
366 studies on other cavefishes, suggest that loss of pigmentation was caused by various
367 kinds of coding region loss-of-function mutation along with widespread
368 transcriptional changes, resulting from extended evolutionary time as a cave-dwelling
369 form.

370 *Oreonectes* fishes could be chosen as representatives of three key nodes of the
371 evolutionary process from a surface-dwelling lifestyle, to semi-cave dwelling, and
372 finally to a permanent cave existence. Almost all *Oreonectes* species show some cave-
373 related traits, such as part or complete eye degeneration and leucism, which makes
374 this genus a good study system of micro-evolution. Therefore, additional genome
375 information of *Oreonectes* species (*de novo* genome assembly, e.g.) ought to be an
376 indispensable resource for a more comprehensive understanding of the adaptive
377 evolutionary mechanism of cavefishes.

378

Declarations

Acknowledgments

We are further grateful to Kai Gao, Jian Chen, Zhonglong Cen and Gang Li for their help in sampling and laboratory work. We thank Martin Burrows and Wenhua Xiong for assistance in writing and data analysis, and two anonymous reviewers for their helpful feedback.

Funding

This project was supported by the key project of Science-technology basic condition platform from The Key Project of Science and Technology Program of Guizhou Province “Study of bilogical and ecological vales for Fanjingshan World Natural Heritage Nomination of Tongren City (3052 2015 Qiankehe SY)”, The Ministry of Science and Technology of the People’s Republic of China (Grant No. 2005DKA21402), Provincial Governor Fund of Guizhou (Qian-Sheng-Zhuan-He-Zi (2010) 14) project and the Youth Innovation Promotion Association of Chinese Academy of Sciences (2011077).

Authors’ contributions

Zhijin Liu, Huamei Wen and Fang Dong performed the experiments, analysed the data, wrote the paper, and participated in the design of the study. Frank Hailer, Fanglei Shi, Zuomin Yang, Tao Liu and Ling Han participated in data analysis, interpretation, and drafting the manuscript. Jiang Zhou and Yibo Hu conceived the study, participated in its design and coordination, and helped to draft the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Our experimental procedures complied with the current laws on animal welfare and research in China, and were specifically approved by the Animal Research Ethics Committee of the Institute of Zoology, Chinese Academy of Sciences.

Competing interest

The authors declare that they have no competing interests.

Consent for publication

All authors gave final approval for publication.

422 **Table 1.** Summary of sample information.

Species	Museum catalog number & GenBank accession number	Sampling location
O. daqikongensis	GZNU20110128002-2011012806	N 25°17′05.1″
	GZNU20151225001-20151225007	E 107°44′54.3″
	mtDNA (NC_039131, GZNU20151225002)	
O. jiarongensis	GZNU20120128001-20120128006	N 25º27′58.82″
	GZNU20151221001-20151221007	E 108º06′47.49″
	mtDNA (KU987437, GZNU20151221001)	
O.dongliangensis	GZNU20160815012-20160815017	N 25°19′48.95″
	mtDNA (KY626008, GZNU20160815012)	E 108°01′18.53″
O. shuilongensis	GNUG20090723002-20090723003	N 25°50′41.87″
	GNUG20160627001-20160627008	E 107°52′57.76″
	mtDNA (KF640641, GZNU20160627001)	

423

424

425

426

427

428

429

430

431

432

433

434 **Table 2.** Primers used for the mtDNA and *MC1R* gene amplifications.

Gene	Primers	Sequence (5' to 3')	Amplicon length
mtDNA	L14724	GAC TTG AAA AAC CAC CGT TG	1192 bp
	H15915	CTC CGA TCT CCG GAT TAC AAG AC	
	L11008	TTG ACT ACC CAA AGC CCA	1249 bp
	H12256	CTT AGA GGG CAA TAG GTG TAA	
	L12166	TGA CAC TGA ATA AAT ACA GCC CT	1022 bp
	H13187	TGA CAC TGA ATA AAT ACA GCC CT	
	L13003	GGT TCC ATT ATT CAC AGC CT	1270 bp
	H14272	TAG GGT TAG TTG CTG TGG C	
	L14080	AAT GGC TCA GCA GCT AAA G	1355 bp
	H15431	TAA TAA ATG GGT GTT CTA CTG G	
MC1R	MC1R-F	GAATATCAGAGGTGTGCTGAAGC	971-1004 bp
	MC1R-R	TCCTTGAGAGTCTTGCGCAG	

435

436

437 **Table 3.** Genetic distances based on P-distance method among *Oreonectes* species.

O. daqikongensis				
O. dongliangensis	0.103			
O. jiarongensis	0.101	0.010		
O. shuilongensis	0.106	0.094	0.091	
O. platycephalus	0.149	0.148	0.147	0.142

438

References

Ayumi T, Hiroaki Y, Jun Y, Cock van O, Masakado K. 2011. The MC1R gene in the guppy (*Poecilia reticulata*): Genotypic and phenotypic polymorphisms. *BMC Research Notes* **4**:31.

Barsh G, Cotsarelis G. 2007. How Hair Gets Its Pigment. *Cell* **130**: 779–781.

Bilandžija H, Ma L, Parkhurst A, Jeffeary WR. 2013. A potential benefit of leucism in Astyanax cavefish: down-regulation of the Oca2 gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. *PLoS One* **8**: e80823.

Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. *Trends in Genetics* **26**: 400–405.

Deng HQ, Wen HM, Xiao N, Zhou J. 2016. A new blind species of the cave genus *Oreonectes* from Guizhou, China (Nemacheilinae). *ZooKeys* **637**: 47–59.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution **29**: 1969–1973.

Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008. Protein structure modeling with MODELLER. *Methods in Molecular Biology* **426**: 145–159.

Frickhinger KA. 1991. Fossilien atlas fishes. Mergus, Melle, Germany. 1088 pp.

Günther A. 1868. Catalogue of the Fishes in the British Museum. Volume seventh .London: Trustees of the British Museum, 512.

Gross JB, Borowsky R, Tabin CJ. 2009. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLoS Genetics* **5**: e1000326.

Gross JB, Wilkens H. 2013. Leucism in phylogenetically and geographically distinct populations of *Astyanax* cavefish arises through the same loss-of-function Oca2 allele. *Heredity* **111**:122–130.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adicoins X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma FD, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Heled J, Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst Biol* **61**: 138-149.

Jeffery WR, Ma L, Parkhurst A. Bilandžija H. 2015. Pigment regression and leucism in Astyanax cavefish. In: Keene AC, Yoshizawa M, McGaugh SE (eds) *Biology and evolution of the Mexican cavefish*. Amsterdam: Elsevier, 155–171.

Johnson JA, Ambers AD, Burnham KK. 2012. Genetics of plumage color in the Gyrfalcon (Falco rusticolus): analysis of the melanocortin-1 receptor gene. *The Journal of Heredity* **103**: 315–321.

Kaelin CB, Barsh GS. 2013. Genetics of Pigmentation in Dogs and Cats. *Annual Review Animal Bioscience* **1**: 125–156.

Kumar CM, Sathisha UV, Dharmesh S, Rao AAG, Singh SA. 2011. Interaction of sesamol (3, 4-methylenedioxyphenol) with tyrosinase and its effect on melanin synthesis. *Biochimie* **93**: 562–569.

Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology* **10**: R83.

Li WZ, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics* **11**: 7.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323–338.

Machado L, Šmíd J, Mazuch T, Sindaco R, Al Shukaili AS, Carranza S. 2018. Systematics of the Saharo- Arabian clade of the Palearctic naked- toed geckos with the description of a new species of Tropiocolotes endemic to Oman. *Journal of Zoological Systematics and Evolutionary Research* **56**: 1–20.

Majerus ME, Mundy NI. 2003. Mammalian melanism: natural selection in black and white. *Trends in Genetics* **19**: 585–588.

Maytin AK, Davies SW, Smith GE, Mullen SP and Buston PM. 2018. *De novo* Transcriptome Assembly of the Clown Anemonefish (*Amphiprion percula*): A New Resource to Study the Evolution of Fish Color. *Frontiers in Marine Science* **5**:284.

Mendes J, Salvi D, Harris DJ, Els J, Carranza S. 2018. Hidden in the Arabian Mountains: Multilocus phylogeny reveals cryptic diversity in the endemic Omanosaura lizards. *Journal of Zoological Systematics and Evolutionary Research* **56**:395–407.

McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery WR, Keene A, Ma L, Minx P, Murphy D, E.Q'Quin K, Rétaux S, Rohner N, Searle SMJ, Stahl BA, Tabin C, Volff JN, Yoshizawa M, Warren WC. 2014. The cavefish genome reveals candidate genes for eye loss. *Nature Communications* **5**: 5307.

Mundy NI, Badcock NS, Hart T, Scribner K, Jansen K, Nadeau NJ. 2004. Conserved genetic basis of a quantitative trait involved in mate choice. *Science* **303**: 1870–1873.

Mundy NI. 2005. A window on the genetics of evolution: *MC1R* and plumage colouration in birds. *Proceedings of the Royal Society of London Series B-Biological Sciences* **272**: 1633–1640.

Nachman MW, Hoekstra HE, D'Agostino SL. 2003. The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences of The United State of America* **100**: 5268–5273.

Nishimura EK. 2011. Melanocyte stem cells: a melanocyte reservoir in hair follicles for hair and skin pigmentation. *Pigment Cell Melanoma Res* **24**: 401–410.

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of leucism. *Nature Genetics* **38**: 107–111.

Posada D, Crandall, KA. 1998. MODELTEST: Testing the model of DNA

substitution. *Bioinformatics* **14**: 817–818.

Ritland K, Newton C, Marshall HD. 2001. Inheritance and population structure of the white-phased "Kermode" black bear. *Current Biology* **11**: 1468–1472.

Rui M, Chen X, Yang J, Richard L. 2012. Phylogenetic relationships of the genus Homatula(Cypriniformes: Nemacheilidae), with special reference to the biogeographic history around the Yunnan-Guizhou plateau. *Zootaxa* **3586**: 78–94.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Stahl BA, Gross JB. 2015. Alterations in *Mc1r* gene expression are associated with regressive pigmentation in *Astyanax* cavefish. *Development Genes and Evolution* **225**: 367–375.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Stern DL. 2013. The genetic causes of convergent evolution. *Nature Reviews Genetics* **14**: 751–764.

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* **17**: 239–250.

Sudhir K, Glen S, Koichiro T. 2016. MEGA7: molecular evolutionary genetics analysis 7.0 version for bigger datasets. *Molecular Biology and Evolution* **33**:1870-1874.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876–4882.

Toyofuku K, Valencia JC, Kushimoto T, Costin GE, Virador VM, Vieira WD, Ferrans VJ, Hearing VJ. 2002. The etiology of oculocutaneous leucism (OCA) type II: the pink protein modulates the processing and transport of tyrosinase. *Pigment Cell Research* **15**: 217–224.

Wang Y, Shen YJ, Feng CG, Zhao K, Song ZB, Zhang YP, Yang LD, He SP. 2016. Mitogenomic perspectives on the origin of Tibetan loaches and their adaptation to high altitude. *Scientific Reports* **6**: 29690.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, 35: 543-548.

Yang J, Chen XL, Bai J, Fang DM, Qiu Y, Jiang WS, Bian C, Lu J, He SY, Pan XF, Zhang YL, Wang XA, You XX, Wang YS, Sun Y, Mao DQ, Liu Y, Fan GY, Zhang H, Chen XY, Zhang XH, Zheng LP, Wang JT, Cheng L, Chen JM, Ruan ZQ, Li J, Yu H, Peng C, Ma XY, Xu JM, He Y, Xu ZF, Xu P, Wang J, Yang HM, Wang J, Whitten T, Xu X, Shi Q. 2016. The Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC Biology* **14**: 1.

Zhan XJ, Dixon A, Fox NC, Bruford MW. 2012. Missense SNP of the MC1R gene is associated with plumage variation in the Gyrfalcon (*Falco rusticolu*s). *Animal Genetic*s **43**: 460–462.

Zhao HB, Yang JR, Xu HL, Zhang JZ. 2010. Pseudogenization of the Umami Taste Receptor Gene Tas1r1 in the Giant Panda Coincided with its Dietary Switch to Bamboo. *Molecular Biology and Evolution* **27:** 2669–2673.

Zhou JJ. 1992. A new Cobitis from the middle Miocene of Shanwang, Shandong. *Vertebrata PalAsiatica* 30: 71-76.

Figure legends

Figure 1. (a) Map of sampling locations: 1. *O. shuilongensis,* 2. *O. jiarongensis*, 3. *O. dongliangensis*, and 4. *O. daqikongensis.* (b) Photo of *O. daqikongensis*, and (c) *O. jiarongensis*. (d) Phylogeny and divergence times among loaches (Nemacheilidae). Red dots denote the utilized calibration points (see Methods). Blue bars show the 95% posterior credibility intervals of divergence times. Number above on each node indicates the Bayesian posterior probability. Red branches indicate the the genus *Oreonectes*. (e) Schematic of MC1R receptor structure and ORF-disrupting mutations of *Mc1r* genes. The first line shows the disrupted sequence of *O. daqikongensis*, lines below represent intact sequences of *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis*. The codon containing an ORF-disrupting mutation (marked red and underlined) is indicated by a box.

Figure 2. Two-dimensional model of MC1R of (a) *O. jiarongensis* and (b) *O. daqikongensis*. (c) Three-dimensional structure of MC1R of *O. jiarongensis,* and (d) and *O. daqikongensis*.

Figure 3. (a) Length distribution of assembled unigenes in *O. jiarongensis* transcriptome. (b) Gene ontology classifications of unigenes from the *O. jiarongensis* transcriptome.

Figure 4. (a) Differentially expressed genes for *O. daqikongensis* skin tissue vs. *O. jiarongensis* skin tissue. Red dots: up-regulated genes: green: down-regulated genes. (b) Schematic mechanisms of signaling pathways for melanogenesis. Down- regulated genes in *O. daqikongensis* skin are shown in red. (c) Expression levels of melanogenesis-related down-regulated genes of *O. daqikongensis* and *O. jiarongensis* measured in FPKM, expression values are mean ± SD of at least three independent experiments.